Estimating the transient climate response from observed warming

Estimating the Transient Climate Response from Observed Warming

ANDREW SCHURER AND GABI HEGERL

School of GeoSciences, University of Edinburgh, Edinburgh, United Kingdom

AURÉLIEN RIBES

CNRM-GAME, Meteo France, CNRS, Toulouse, France

DEBBIE POLSON

School of GeoSciences, University of Edinburgh, Edinburgh, United Kingdom

COLIN MORICE

Met Office Hadley Centre, Exeter, United Kingdom

SIMON TETT

School of GeoSciences, University of Edinburgh, Edinburgh, United Kingdom

(Manuscript received 23 October 2017, in final form 31 July 2018)

ABSTRACT

The transient climate response (TCR) quantifies the warming expected during a transient doubling of greenhouse gas concentrations in the atmosphere. Many previous studies quantifying the observed historic response to greenhouse gases, and with it the TCR, use multimodel mean fingerprints and found reasonably constrained values, which contributed to the IPCC estimated (>66%) range from 1° to 2.5°C. Here, it is shown that while the multimodel mean fingerprint is statistically more powerful than any individual model's fingerprint, it does lead to overconfident results when applied to synthetic data, if model uncertainty is neglected. Here, a Bayesian method is used that estimates TCR, accounting for climate model and observational uncertainty with indices of global temperature that aim at constraining the aerosol contribution to the historical record better. Model uncertainty in the aerosol response was found to be large. Nevertheless, an overall TCR estimate of 0.4°–3.1°C (>90%) was calculated from the historical record, which reduces to 1.0°–2.6°C when using prior information that rules out negative TCR values and model misestimates of more than a factor of 3, and to 1.2°–2.4°C when using the multimodel mean fingerprints with a variance correction. Modeled temperature, like in the observations, is calculated as a blend of sea surface and air temperatures.

1. Introduction

The detection and attribution of the causes of climate change seek to disentangle the changes caused by known drivers, like greenhouse gas (GHG) emissions, anthropogenic aerosols, and natural forcings such as volcanic eruptions from that of internal variability and have been used extensively to determine the past and predicted anthropogenic contribution to global warming (see, e.g., Bindoff et al. 2013, and references therein). The techniques commonly used compare the spatio-temporal temperature pattern in observations to that produced by models and employ sophisticated statistical analysis to separate the forced from unforced variability. Typically these rely on optimized linear regression, a technique first proposed by Hasselmann (1993), that was further developed by Hegerl et al. (1997), Allen and Tett (1999), Allen and Stott (2003), and Ribes et al. (2013) to

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-17-0717.s1.

Corresponding author: Andrew Schurer, a.schurer@ed.ac.uk

DOI: 10.1175/JCLI-D-17-0717.1

determine the combination of the responses to different forcings, often referred to as their fingerprints, which best fit the observations. In practice, the attributable response to a particular forcing is estimated by calculating scaling factors for each of the forcings. If the scaling factor for a forcing is found to be significantly greater than zero, the effect of the forcing has been detected.

Through the use of these methods, it was found that the attributable warming from GHGs could be consistently constrained using several different models (Stott et al. 2006). This is important since the effect of past GHG emissions on climate can be used to estimate the likely impact of future GHG increases. One way this can be done is by using detection and attribution results to constrain the transient climate response (TCR) or closely related transient climate response to cumulative carbon emissions (TCRE). The TCR is formally defined as the warming at the time of CO_2 doubling in a climate simulation where the CO_2 concentration is gradually increased by 1% yr^{-1} (Hegerl et al. 2007). The TCR has been found to be a generic property of the climate system that can be used to determine the global climate temperature response to any gradual increase in forcing (Bindoff et al. 2013). Previous studies have found that a probabilistic estimate of TCR can be calculated from the scaling factors derived from detection and attribution results (Frame et al. 2006). The TCRE combines the transient warming response with information about the carbon cycle and can be used to estimate the global warming response to cumulative emissions of CO_2. It can be estimated from observations by dividing the attributable warming due to CO_2 by historical cumulative carbon emissions (Allen et al. 2009; Matthews et al. 2009; Gillett et al. 2013). Alternatively, the scaling factors calculated by detection and attribution studies can be used to scale model simulations of future projections helping to provide a constraint to future predictions (Kettleborough et al. 2007).

Recent detection and attribution studies carried out using the large number of simulations made available by phase 5 of the Coupled Model Intercomparison Project (CMIP5; Taylor et al. 2012) have been able to constrain attributable warming due to GHGs using some models' fingerprints, while results based on other models give very wide or unconstrained values (Gillett et al. 2013; Jones et al. 2013, 2016; Ribes and Terray 2013). This has raised the question of whether detection and attribution can be reliably used to constrain predicted warming. As investigated in Jones et al. (2016), the problem arises when attempting to separate the response to anthropogenic aerosol forcing from that of GHGs, which tend to show similar, but opposite patterns leading to degenerate results when attempting to separate their effect in observations. This is because GHGs cause warming and aerosols cooling, with similar spatial patterns and time scales (Wilcox et al. 2013; Xie et al. 2013), although the level of similarity varies between models. The choice of a metric to best separate their effects is complicated by the differences between the responses to anthropogenic aerosols in different models. One proposed solution is to use average patterns across all models (which will subsequently be called the multimodel mean). This has been found to give constrained attributable GHG warming (Bindoff et al. 2013), and it is results using the multimodel mean that have been used in a number of studies (e.g., Jones et al. 2013; Gillett et al. 2013) to estimate a value for TCR. These formed one strand of evidence which allowed the Intergovernmental Panel on Climate Change (IPCC) to estimate a relatively wide likely ($p > 0.66$) range of 1°–2.5°C for the TCR (Collins et al. 2013), using expert opinion to aggregate across multiple lines of evidence and account for further uncertainties. The use of fingerprints taken from a multimodel mean is supported by studies showing that multimodel average results often outperform any individual model (Knutti et al. 2010). In addition, by averaging over all available simulations, the internal variability on the model fingerprints is substantially reduced, giving better-constrained TCR estimates as a consequence (Ribes et al. 2015). One problem relying on the multimodel mean, though, is that it does not account for uncertainty in the pattern of response to forcing, which can vary between models.

Huntingford et al. (2006) proposed using an errors in variables (EIV) approach to explicitly account for model uncertainty, an approach that has also been used by Gillett et al. (2013) to calculate TCR. However, in these approaches the statistical inference is much more complicated (e.g., maximum likelihood estimates are not explicit), and there are remaining problems in the uncertainty analysis and calculation of the model uncertainty covariance related to incomplete sampling of the true model error covariance (Schnur and Hasselmann 2005; Hannart et al. 2014) from a limited and biased sample of models (Knutti et al. 2010). Recently, Ribes et al. (2017) introduced a new statistical approach to the detection and attribution problem. Instead of a regression-like approach, they used an additive decomposition technique to determine the most likely contribution from different forcings. Unlike traditional detection and attribution, this approach explicitly accounts for modeling uncertainty, which is estimated using the ''models are statistically indistinguishable from the truth'' paradigm. This approach has been used to calculate the contribution of different forcings to recent temperature rise. This method could also be used to determine a TCR estimate.

However, the paradigm used in Ribes et al. (2017) would imply that the TCR distribution is roughly restricted to the modeled range. This is equivalent to assuming that we have fairly constrained prior information about the TCR range coming from model simulations.

Here we will consider wider (i.e., less informative) priors, in the standard detection and attribution framework of a regression-like approach, in order to estimate TCR. We consider the effect of observational uncertainty by the use of an ensemble of instrumental observations, and model uncertainty by calculating the results from seven different models, and combine the results by integrating over the probability density functions arising from individual realizations of model response [as was done in Hegerl et al. (2006) for many possible reconstructions of past climate]. This framework allows the use of prior information, which we find helps separate the response to the different forcings, leading to a more constrained estimate of TCR, and prevents unphysical scaling factors. We represent the spatial pattern by the global mean, hemispheric contrast, seasonal difference, and land–sea contrast. These are used to help separate the GHG warming from the aerosol cooling. By using relatively simple spatial information we also remove the need for projection into a low-dimensional space, often of empirical orthogonal functions (EOFs) of internal variability, which reduces the complexity and removes one possible source of uncertainty (Ribes et al. 2013) in addition to avoiding loss of signal due to inability to capture it in a low-dimensional noise EOF space (Hegerl et al. 1997).

In section 2 we present the modeled and observed temperature data and discuss our choice of fingerprints. Section 3 outlines the analysis framework and our statistical method. Section 4 describes tests carried out to evaluate the method using synthetic data from models in the place of observations to estimate known TCR values. These include a ''perfect model study,'' where the model used as observation is taken from the same model as that used for the analysis, and an ''imperfect model study,'' where the model used as observation is taken from a different model. In section 5, TCR estimates from observed temperatures are shown and the implications discussed, followed by conclusions (section 6).

2. Temperature data

a. Observations and models

The analysis is carried out, for each model, over the period 1863–2012 on decadally averaged temperature data (leading to 15 independent time values). We analyze a range of spatial and seasonal temperature fields, combining the global mean temperature with the hemispheric temperature difference as well as seasonal and land–ocean temperature contrasts.

For the observations we use the HadCRUT4 dataset (version 4.5.0.0; Morice et al. 2012). This is a gridded blend of the CRU Air Temperature Anomalies, version 4 (CRUTEM4) land surface air temperature dataset and the HadSST3 sea surface temperature (SST) dataset, where for each location anomalies are calculated with respect to 1961–90. This dataset only contains temperature values in grid cells where there are observations; consequently, the earlier decadal means, when observations were sparse, will be based on much fewer data points than the final decades (decadal means have under 20% total coverage for 1863–72, rising to about 80% coverage in the Northern Hemisphere and 60% in the Southern Hemisphere by the end of the twentieth century). In this study we calculate an ensemble of 100 possible realizations of temperature based on the HadCRUT4 dataset (Morice et al. 2012), with additional uncertainty information provided in HadCRUT4 encoded into the ensemble. The HadCRUT4 ensemble time series, which sample uncertainty associated with systematic changes in observing practices, are augmented by additionally sampling from the measurement uncertainty and within gridcell sampling uncertainty terms of the Morice et al. (2012) uncertainty model. In the HadCRUT4 ensemble dataset this information is included as a separate uncertainty term that is not encoded into the ensemble time series (for more details, see the online supplementary material). Uncertainty relating to incomplete coverage of the globe by the HadCRUT4 grid is accounted for by masking the model fields to match the available coverage in HadCRUT4, as has been the methodology in previous detection and attribution studies (see Bindoff et al. 2013, and references therein), and is not encoded into the ensemble time series used in this study.

We note that the observational ensemble only pertains to the Morice et al. (2012) method. Other near-surface temperature datasets use different methods to homogenize and bias adjust observational data, and uncertainties associated with these different methods are not explored. In particular, different methods for bias adjustment of sea surface temperature measurements have an impact over decadal time scales (Kent et al. 2017) that are not explored in this study. By repeating our analysis with each sample (weighting each equally), the analysis results should span the full HadCRUT4 uncertainty range.

For the model fingerprints, we use historical CMIP5 simulations from 1863 to 2012. We apply the methodology

TABLE 1. CMIP5 models used in the analysis. Bold values are the number of ensemble members for the individual models used in the main analysis. Numbers in parentheses are the number of model ensemble members contributing to the multimodel mean. Note NorESM1-M was only used for the multimodel mean. Where results using individual models are shown, they will be plotted using the colors found in the last column. TCR values calculated from 1PCT simulations.

Model	Number of historical simulations	Number of historicalGHG simulations	Number of historicalNAT simulations	TCR value (°C)	Color in figures
CanESM2	**5** (5)	**5** (5)	**5** (5)	2.4	Light green
CNRM-CM5	**10** (6)	**6** (6)	**6** (6)	2.1	Red
CSIRO Mk3.6.0	**10** (5)	**5** (5)	**5** (5)	1.8	Teal
GISS-E2-H (p1)	**5** (5)	**5** (5)	**5** (5)	1.7	Purple
GISS-E2-R (p1)	**6** (5)	**5** (5)	**5** (5)	1.5	Orange
HadGEM2-ES	**4** (4)	**4** (4)	**4** (4)	2.5	Pink
IPSL-CM5A-LR	**4** (3)	**3** (3)	**3** (3)	2.0	Blue
NorESM1-M	(1)	(1)	(1)	1.4	—
Multimodel mean	(34)	(34)	(34)	2.0	Black

outlined in Cowtan et al. (2015), to combine climate model surface air temperature (SAT) and SST together to produce blended surface temperatures that can be directly compared to HadCRUT4. This is important as the commonly used approach of just using model SATs over both land and ocean was found to substantially underestimate climate sensitivity (Richardson et al. 2016). This method masks the model data to where there are observations in HadCRUT4, using the SAT field for land-only and ice-covered grid boxes and the SST field for ice-free ocean regions. Similar to the observations, anomalies in each grid cell are calculated with respect to 1961–90. For details, see Cowtan et al. (2015). Using the blended temperature is fairly rare in a detection and attribution of this type, with previous studies (Gillett et al. 2013; Jones et al. 2013; Ribes and Terray 2013) comparing observed changes in temperature with the model SAT field. The implications of this will be discussed later in the paper.

We restrict our analysis to only include models that simulate both a direct radiative aerosol effect and an indirect aerosol effect (which includes interactions with clouds), and therefore represent the anthropogenic aerosol forcing in a physically more realistic manner. It has also been found that this will better represent observed temperature changes (Wilcox et al. 2013). In common with previous studies (e.g., Gillett et al. 2013; Jones et al. 2013), we use a three signal analysis with fingerprints given by the historicalNAT (natural forcings, such as solar and volcanic only), historicalGHG (GHG forcings only), and historical (all forcings) CMIP5 simulations to calculate contributions by different combinations of forcings. Ideally, we would not restrict our analysis to only three fingerprints but decompose the temperature change into more factors. This, however, has been found to lead to degenerate results (Tett et al. 2002). We use ensemble means for all models which cover the analysis period (1863–2012) with at least three ensemble members for each experiment (see Table 1 for details of the models used). The historical simulations only cover the period up to 2005. To extend the analysis to 2012, we preferentially use the CMIP5 historicalExt experiments, or if not available, RCP4.5 experiments.

The 1% yr^{-1} (1PCT) CO_2 simulations are used to estimate a model's TCR value. As in Gillett et al. (2013), we calculate the TCR value using the global mean temperature anomaly in the 1PCT simulation, relative to the corresponding control simulation, as the 20-yr mean temperature centered on the year 70. We estimate a distribution of possible values, given internal variability, using the model's piControl simulations.

Multimodel mean fingerprints are calculated from all models with simulations available from each of the three different experiments and are constructed to have the same mix of model simulations across all experiments. This means, for example, that if a model has six historical and historicalGHG but only four historicalNAT simulations, then only the first four ensemble members from each experiment will be used in the multimodel mean, which is calculated as the mean over all simulations. The multimodel mean TCR value (in Table 1) is calculated from the multimodel mean of the respective 1PCT simulations for the different model simulations used taking into account the relative number of simulations each model contributes to the multimodel mean. Calculating a mean over the available simulations and not the mean over all models will result in a greater signal-to-noise ratio but will lead to some models being given greater weight.

For the imperfect model study all the historical simulations listed in Table 1 are used as pseudo observations to act as a target for our analysis. In addition, further models, which do not provide individually forced simulations but do include both the direct and indirect

TABLE 2. Additional CMIP5 models used in the imperfect model analysis. Total number of simulations in the final row includes historical simulations used in the main analysis (see Table 1). TCR values from 1PCT simulations.

Model	Number of historical simulations	TCR value (°C)
ACCESS1.0	1	2.0
ACCESS1.3	1	1.7
CESM1(CAM5)	3	2.3
GISS-E2-H (p3)	5	1.7
GISS-E2-R (p3)	6	1.5
GFDL-CM3	1	2.0
HadCM3	10	2.0
IPSL-CM5A-MR	1	2.0
MIROC5	1	1.5
MIROC-ESM	1	2.2
MRI-CGCM3	1	1.6
NorESM1-ME	1	1.6
Total simulations for imperfect model study (including models from Table 1)	77	—

anthropogenic aerosol effects, are included, bringing the total number of simulations used in the imperfect model analysis to 77 (see Table 2).

To derive samples of internal variability, we use 150-yr segments from all piControl simulations in the CMIP5 archive (see supplementary material for details of models used) where each sample is a blend of SST and SATs masked to the HadCRUT4 dataset in the same way as with the forced experiments; consequently, our samples of internal variability will account for the change in observational coverage. In total, this gives us 140 independent chunks of 150 years in length. To increase the number of degrees of freedom of the internal variability estimate, we take one sample every 25 years, resulting in a total of 673 samples with an estimated 210 degrees of freedom (Allen and Tett 1999).

b. *Choice of fingerprints*

Fingerprints of change are calculated for both the observations and the model data, as a blend of SATs and SSTs (see previous section). The resulting decadal time series (see Fig. 1), are formed of 15 time values and have anomalies calculated with respect to the full period (i.e., the mean of all 15 points is subtracted).

Figure 1 compares the response to all forcings in individual models and the multimodel mean with that of the observations. It also shows the contribution from different combinations of forcings. For illustrative purposes only the response to ''other'' anthropogenic forcings is calculated from the historical simulation from which the historicalGHG and historicalNAT ensemble means are subtracted [note that this time series is not directly used in the analysis—see Eqs. (8)–(10)]. The resultant ''other'' anthropogenic forcing time series will be predominantly driven by anthropogenic aerosols (Myhre et al. 2013).

Our first choice of diagnostic is the global mean temperature. This is calculated as the mean temperature of the Northern Hemisphere (NH) and Southern Hemisphere (SH), as used in Morice et al. (2012). The response to forcing in global mean surface temperature has been shown to be a good choice for detecting climate change signals due to a large signal to noise ratio (Hegerl et al. 1997; Hegerl and North 1997; see also Ribes et al. 2013). The strong degeneracy between the response to GHG forcing and aerosol forcing, reported by previous studies (Wilcox et al. 2013; Xie et al. 2013), is clear though in the global mean temperature if the blue and red lines are compared in Fig. 1b, with the aerosols causing cooling while the GHGs cause warming. Our choice of further spatial and temporal diagnostics is influenced by this need to separate the GHG and aerosol forcings, so in addition to the global mean we select other diagnostics, which have been suggested in the literature.

Aerosols are predominantly emitted over NH land, so they could be expected to have a larger effect over these areas (see Ming and Ramaswamy 2009; Shindell 2014; Stevens 2015; Wang et al. 2016). Consequently, we choose to look at the hemispheric difference and the land–ocean contrast. It has also been shown that the modeled response to anthropogenic aerosols could have a strong seasonal signal, with more cooling during summer months (Hegerl et al. 1997; Tett et al. 2007), so this will be investigated through use of a seasonal contrast, which will be calculated as the difference in mean temperature in October–March compared to April–September in the NH.

The hemispheric difference appears to be a good choice to separate GHGs and aerosols. Although the aerosols preferentially cool the NH with respect to the SH while the GHGs preferentially warm the NH, their

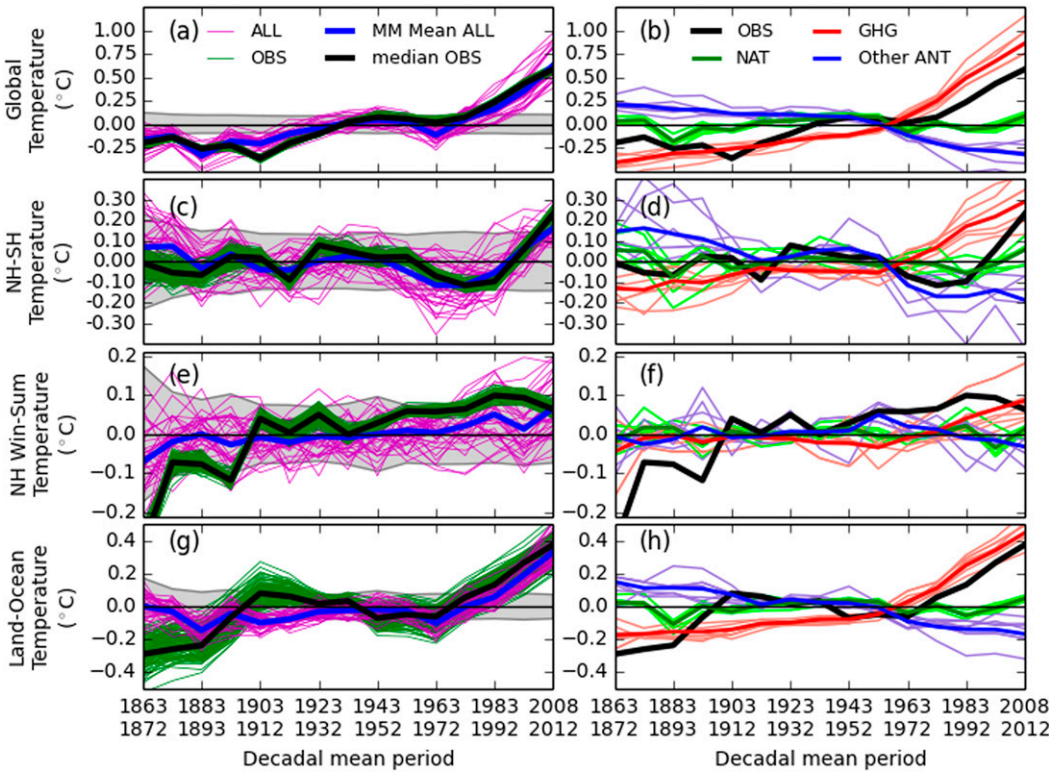

FIG. 1. Decadal mean temperature for observations and models for four different diagnostics for climate change: (a),(b) Global temperature, (c),(d) hemispheric temperature difference, (e),(f) seasonal contrast (NH winter temperature mean subtracted by NH summer mean), and (g),(h) land–ocean contrast (global land mean temperature subtracted by global ocean temperature). (left) Individual observation ensemble members (green) and median (black). Multimodel mean all forced historical simulations (blue) and individual model simulations (purple). (right) GHG-only simulation (historicalGHG; red), natural forcings only (historicalNat; green), and other-ANT (blue), which is plotted from the historical experiment mean after subtraction of the historicalGHG and historicalNat mean. Bold lines are for multimodel means, and thin lines are for individual model ensemble means. The gray range in the left panels shows two standard deviations of piControl simulations.

different temporal histories lead to responses that appear to break this degeneracy. Although there is a clear difference in the response to the forcing in each model, in general the aerosols have a larger effect up to the 1970s, while after the 1980s the situation reverses with the GHG effect dominating. The resultant temperature response to the combined forcings (Fig. 1c) looks qualitatively consistent with that observed, suggesting that much of the observed change in hemispheric contrast is forced.

For the seasonal contrast (Figs. 1e,f), the response to GHGs shows a positive trend, indicating that they cause more warming in winter months than summer months in the NH, which is consistent with the observed change (Fig. 1e). The seasonal response to aerosols is very model specific with aerosol forcing, in some models causing positive and some negative trends. Whether including the seasonal contrast in the regression provides a better constraint therefore may prove to be model dependent.

In the response in the land–ocean contrast (Figs. 1g,h), GHG forcing warms the land more than the ocean, while aerosols cool the land more than the ocean. Thus, the response to this diagnostic shows a similar degeneracy to that seen in the global mean temperature (Fig. 1b) and may consequently be of limited use to disentangle the different forcing responses. In addition, the large observational uncertainty and clear discrepancy between the observed and modeled land–ocean contrast early in the record could prove problematic.

The uncertainty in the observations in Fig. 1 is represented by the ensemble spread (see section 2). It is small for the global mean, larger for the hemispheric and seasonal differences, and larger still for the land–ocean contrast. The observations lie within the model range for the first three panels, except in the global mean, where there is a discrepancy in the 1910s (although statistically, a short period of discrepancy is

unsurprising). For the land–ocean contrast, many of the observational ensemble members are outside the model range for much of the analysis period, particularly during the early part, again supporting an omission of this diagnostic.

To test which of these diagnostics is best to disentangle the forced responses, we will carry out our analysis on a range of different choices for spatial and temporal means. Specifically, we will analyze four different choices: the global mean temperature; global mean temperature combined with hemispheric temperature difference; global mean temperature, hemispheric difference, and seasonal difference; global mean temperature, hemispheric difference, and land–ocean difference.

3. Methods

a. Analysis framework

We start from the standard optimal detection framework (see, e.g., Hasselmann 1993; Allen and Stott 2003; Hannart et al. 2014; Ribes et al. 2013) and assume that observed changes in surface temperature $\mathbf{Y}$ are due to a sum of an externally forced components and internal variability $\boldsymbol{\varepsilon}_0$. The externally forced components can be estimated from a linear combination of i scaled model fingerprints $\mathbf{X}_i$, which also contain internal variability $\boldsymbol{\varepsilon}_i$. Following the formalization in Ribes et al. (2013), the problem can be expressed as

$$\mathbf{Y} = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0, \quad \boldsymbol{\varepsilon}_0 \sim N(0, \boldsymbol{\Sigma}), \tag{1}$$

$$\mathbf{X}_i = \mathbf{X}_i^* + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\Sigma}/n_i), \tag{2}$$

where $\mathbf{X}^*$ is the true (i.e., noise free) model response of the climate system, with columns $\mathbf{X}_i^*$, and $\boldsymbol{\varepsilon}_i$ denoting a random term that is assumed to be entirely due to the internal variability of model fingerprint i. The fact that the fingerprint $\mathbf{X}_i$ is calculated as an ensemble mean (or a multimodel mean) over n_i simulations implies that the variance of $\boldsymbol{\varepsilon}_i$ is $\boldsymbol{\Sigma}/n_i$. This is the approach used in the classical detection analyses, although Allen and Stott (2003), who initially introduced the total least squares (TLS) technique, wrote the statistical model slightly differently.

We wish to use this framework to calculate a likelihood for the β_i values. We first assume that internal variability ($\boldsymbol{\varepsilon}_0$, $\boldsymbol{\varepsilon}_i$) is Gaussian, which is reasonable for long-term and large-scale temperatures due to the central limit theorem. Term $\ell(\mathbf{X}^*, \boldsymbol{\beta})$, the −2 log-likelihood of the parameters $\mathbf{X}^*$ and $\boldsymbol{\beta}$, can then be written as

$$\ell(\mathbf{X}^*, \boldsymbol{\beta}) = C + (\mathbf{Y} - \mathbf{X}^*\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}^*\boldsymbol{\beta}) + \sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{X}_i^*)'\left(\frac{\boldsymbol{\Sigma}}{n_i}\right)^{-1}(\mathbf{X}_i - \mathbf{X}_i^*), \tag{3}$$

which is a function of $\mathbf{X}^*$ and $\boldsymbol{\beta}$, the two parameters in Eqs. (1) and (2), and where C is a constant that does not depend on any of those two parameters.

The conventional TLS estimate minimizes this −2 log-likelihood ℓ. To do that, we minimize ℓ on $\mathbf{X}^*$, for any fixed $\boldsymbol{\beta}$. In that way, we compute the profile likelihood, or concentrated likelihood

$$\ell_c(\boldsymbol{\beta}) = \min_{\mathbf{X}^*} \ell(\mathbf{X}^*, \boldsymbol{\beta}), \tag{4}$$

$$\ell_c(\boldsymbol{\beta}) = C + (\mathbf{X}\boldsymbol{\beta} - \mathbf{Y})'(S\boldsymbol{\Sigma})^{-1}(\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}), \tag{5}$$

where

$$S = \sum_i \frac{\beta_i^2}{n_i} + 1. \tag{6}$$

In our analysis we wish to solve Eqs. (1) and (2) for a linear combination of the historicalNAT (simulations forced by only natural forcings), historicalGHG (simulations forced by only GHG forcings), and historical (all forcings) simulations (see section 2):

$$\mathbf{Y}(t) = \mathbf{X}^*_{\text{historical}}\beta_{\text{historical}} + \mathbf{X}^*_{\text{historicalGHG}}\beta_{\text{historicalGHG}} + \mathbf{X}^*_{\text{historicalNAT}}\beta_{\text{historicalNAT}} + \boldsymbol{\varepsilon}_0. \tag{7}$$

Equation (5) is used to estimate the likelihood of scaling factors β_{histALL}, β_{histGHG}, and β_{histNAT} (where these estimates are denoted by $\hat{\boldsymbol{\beta}}$).

Since the historical simulations contain the response to all known forcings including GHGs and natural forcings, those need to be disentangled to isolate just the total contribution from GHGs, β_{GHG} (which is needed to determine the likelihood function of the TCR). From Eq. (7), assuming linearity, and the relationship $\mathbf{X}_{\text{ALL}}(t) = \mathbf{X}_{\text{otherANT}}(t) + \mathbf{X}_{\text{NAT}}(t) + \mathbf{X}_{\text{GHG}}(t)$, it follows as described in Tett et al. (2002) that

$$\hat{\beta}_{\text{otherANT}} = \hat{\beta}_{\text{historical}}, \tag{8}$$

$$\hat{\beta}_{\text{GHG}} = \hat{\beta}_{\text{historicalGHG}} + \hat{\beta}_{\text{historical}}, \quad \text{and} \tag{9}$$

$$\hat{\beta}_{\text{NAT}} = \hat{\beta}_{\text{historicalNAT}} + \hat{\beta}_{\text{historical}}. \tag{10}$$

Using the likelihoods calculated in Eq. (3) with Eqs. (8)–(10), we can calculate a likelihood of any combination of β_{otherANT}, β_{GHG}, and β_{NAT}.

Using Bayesian inference, prior knowledge about the density function of any combination of $\boldsymbol{\beta}$ values can be

included in our analysis. This prior information $\mathbf{P}(\boldsymbol{\beta})$ is then updated using the evidence from observations from the optimal detection analysis. Following Bayes theory we can calculate a conditional probability of $\boldsymbol{\beta}$, given the observationally constrained estimate, using Eq. (11) (see also Lee et al. 2005; Berliner et al. 2000):

$$\mathbf{P}(\hat{\boldsymbol{\beta}} \,|\, \boldsymbol{\beta}) = \frac{\mathbf{P}(\hat{\boldsymbol{\beta}} \,|\, \boldsymbol{\beta})\mathbf{P}(\boldsymbol{\beta})}{\int \mathbf{P}(\hat{\boldsymbol{\beta}} \,|\, \boldsymbol{\beta})\mathbf{P}(\boldsymbol{\beta})\, d\boldsymbol{\beta}}, \tag{11}$$

where $\mathbf{P}(\hat{\boldsymbol{\beta}} \,|\, \boldsymbol{\beta})$ is our likelihood estimate taken from the optimal detection analysis.

To obtain an estimate of TCR, for an individual model we first need to calculate the likelihood of the scaling factors for GHGs $\mathbf{P}(\beta_{\mathrm{GHG},j} \,|\, \hat{\beta}_j)$ and then multiply the values by each model's actual TCR value $\mathbf{P}(\mathrm{TCR}_j)$ calculated from each model's 1PCT simulations (see section 2). This is possible because TCR approximately scales with the observed GHG warming (Frame et al. 2006), although the spread in this relationship gives rise to additional uncertainty (Gillett et al. 2013). For any model j, the probability density of the amplitude of the GHG scaling factor can therefore be used to calculate probabilities for TCR. To combine these into a single distribution, we calculate a weighted distribution that is integrated across all the different models j, where each model here is given equal weighting; this is equivalent to Bayesian model averaging with a uniform prior and is similar to the approach in Hegerl et al. (2006):

$$\mathbf{P}(\mathrm{TCR} \,|\, \hat{\beta}_{\mathrm{GHG},j}) = \sum_j \frac{1}{n_j} \mathbf{P}(\beta_{\mathrm{GHG},j} \,|\, \hat{\beta}_{\mathrm{GHG},j}) \mathrm{TCR}_j. \tag{12}$$

b. Prior information

Typically in optimal detection studies (e.g., Gillett et al. 2013; Jones et al. 2013; Ribes and Terray 2013), all values of scaling factors are treated as equally likely, including negative scaling factors. This leaves feedbacks to forcings response completely unconstrained and varying between forcings, ignoring some basic physical constraints (Hegerl and Zwiers 2011). Here we use prior information on the scaling factors [$\mathbf{P}(\boldsymbol{\beta})$, Eq. (11)] to treat values closer to the modeled values as more likely, while at the same time giving zero possibility to physically implausible very large and negative values. This is similar to Hegerl and Allen (2002), who constrained the amplitude of the anthropogenic aerosol signal to be positive, improving the attribution of the GHG contribution.

Our choice of the main analysis prior is motivated by three primary considerations:

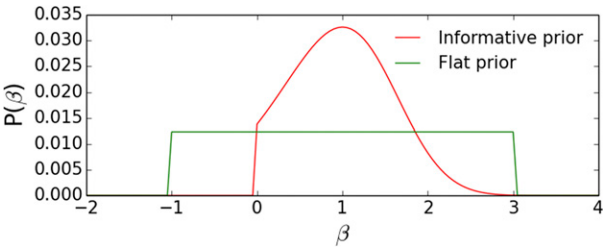

FIG. 2. Choice of priors: informative priors on β_{GHG}, $\beta_{\mathrm{otherANT}}$, and β_{NAT} in red and uninformative priors in green.

1) That the prior had a mode and median value of 1 (i.e., the most likely value is 1, and the model is equally likely to have a scaling factor greater than 1 as less than 1).
2) That the probability of having a scaling factor of 0 is finite, but the probability of negative scaling factors is zero; this rules out physically implausible responses while still allowing for the response to that forcing not being present.
3) That very large scaling factors above a factor of 3 are impossible. This is a strong constraint, but given the model's TCR values (see Table 1), this constraint will only rule out very large scaled TCR values above 4.5°C, which are far outside the range covered in climate models and have already been found to be inconsistent by multiple lines of evidence (Collins et al. 2013; Knutti et al. 2017).

Following these considerations, we have chosen a skew-normal distribution for our prior distribution on β_{GHG}, β_{NAT}, and $\beta_{\mathrm{otherANT}}$ (see Fig. 2). To investigate the sensitivity of our results to the choice of prior, we also repeat all our results using a less informative prior on β_{GHG}, β_{NAT}, and $\beta_{\mathrm{otherANT}}$, which is constant from −1 to 3. The latter will be referred to as uninformative prior in the following, although it already implies some constraint on the scaling factors, with large negative and large positive values excluded. The IPCC report (Myhre et al. 2013) found that there was a very small probability of anthropogenic aerosols having a positive forcing mainly due to black carbon, which would require a negative anthropogenic aerosol scaling factor for the models used here, and this will be allowable under this less informative prior.

c. Analysis framework summary

In summary, to calculate a TCR value, our overall analysis strategy is as follows:

1) For each individual model, calculate the likelihood distribution of scaling factors for a linear combination of the ensemble mean of historical, historicalGHG, and historicalNAT experiments given a particular set of observations [Eq. (5)].

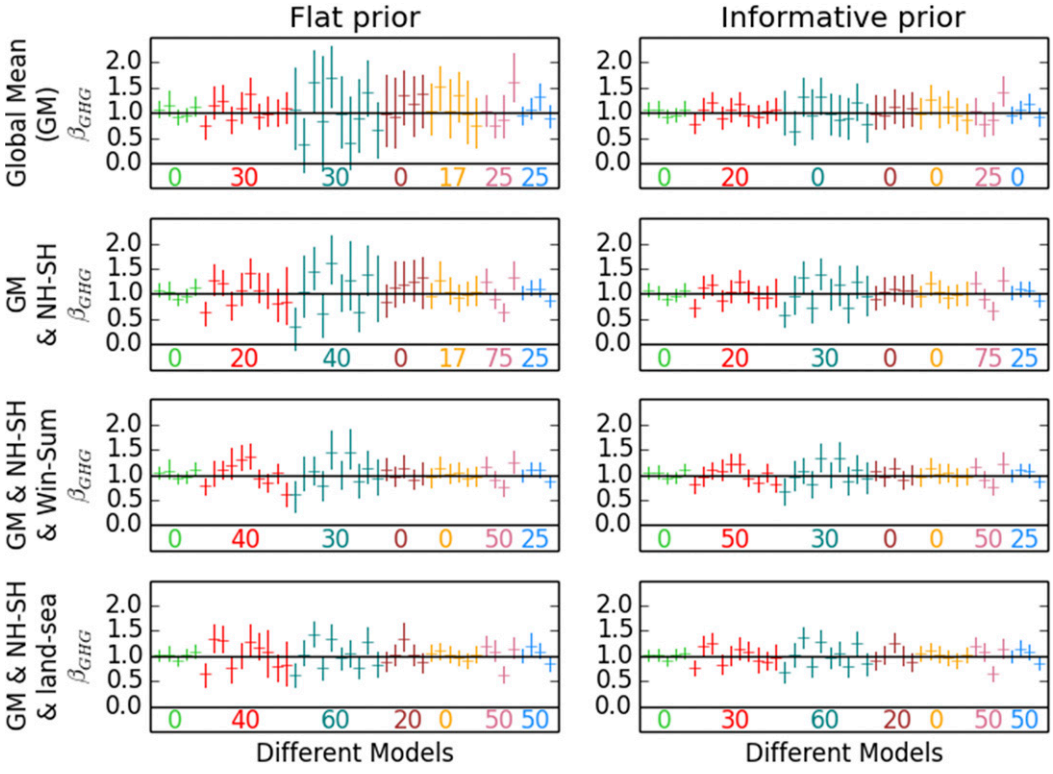

FIG. 3. Perfect model results for different combinations of spatial and temporal diagnostics: 5%–95% and median probability range of the GHG scaling factor β_{GHG}, where a historical simulation has been used as observations and the fingerprints are taken from the same model. Panels show results for (top) global mean temperature; (top middle) global mean temperature combined with hemispheric temperature difference; (bottom middle) global mean temperature, hemispheric difference, and seasonal difference; and (bottom) global mean temperature, hemispheric difference, and land–ocean difference where (left) a noninformative prior has been used and (right) an informative prior has been used. Colors show results for different models (see Table 1). Numbers at the bottom of each panel indicate the percentage of cases where the 5%–95% range does not include 1.

2) Determine a likelihood distribution for every combination of scaling factors $\beta_{\mathrm{otherANT}}$, β_{GHG}, and β_{NAT} individually using Eqs. (8)–(10).
3) Use these calculated likelihoods to update the prior information (Fig. 2) using Eq. (11).
4) Convert the probability distribution $\mathbf{P}(\beta_{\mathrm{GHG}})$, for a specific model, to a probability distribution for TCR by scaling $\mathbf{P}(\beta_{\mathrm{GHG}})$ by the TCR value of that model (not accounting for uncertainty in this relationship).
5) Repeat for all models, including the multimodel mean.
6) Integrate results from individual models into a combined distribution by calculating an average equally weighted distribution [Eq. (12)] and present the results as the 5%–95% and median values of this distribution (plotted in purple in Figs. 4–9).
7) To account for observational uncertainty, repeat analysis using each of the observational ensemble members. Integrate over all distributions and present the results as the 5%–95% and median values of this distribution.

4. Tests of the analysis method

a. *Perfect model study*

To both test that the analysis produces unbiased results and to help determine which of our diagnostics is best at constraining the GHG response, we have conducted a perfect-model study, whereby our analysis setup is used to detect the likelihood of the forcing contribution in an all-forced historical simulation from the same model. Specifically, this means that we use one of the historical simulations as pseudo observation and use all the other historical, historicalNAT, and historicalGHG simulations to determine the likelihood of each range of scaling factors $\boldsymbol{\beta}$. If our statistical model (including the assumption of linearity) is correct, by definition this should give a 5%–95% confidence interval, which includes 1 for every scaling factor in 9 out of 10 cases on average. The results for our choice of four different spatial and temporal diagnostics (Fig. 3) show that we can estimate a reasonable value for β_{GHG} of a

model in this ''perfect model'' setup, although the uncertainties are relatively large. Some models give tighter constraints than others and these uncertainties are reduced as the analysis uses more spatial and seasonal information, suggesting that as the hemispheric difference improves, as expected, the GHG constraint and the inclusion of seasonal and land–ocean contrast further improves the analysis.

It is clear that some models give overconfident results (meaning that the 5%–95% range does not contain the true value of 1 for 10% of the time), with the problem often more acute when the land–ocean contrast is used. One possible explanation for this could be that the covariance matrix used to determine the likelihoods [Eq. (3)] is calculated from segments from control simulations from a large number of different models, so it will not be the true covariance matrix for any one individual model used in the analysis. Consequently, models that have larger internal variability might be expected to have overconfident results and vice versa. To investigate this, we have compared the decadal internal variability of the global mean temperature for each of the models (estimated from their control simulations) and compared it to the distribution of the full set of control simulations used in the analysis (see Fig. S1). As could be expected, the models with less variability than the average of the distribution—CanESM2, GISS-E2-H, and GISS-E2-R—have failure rates similar to or less than the expected 10%. In contrast, the models that are more variable than the average are those which give apparently overconfident results. Ideally, only samples from the same model control simulations should be used, but this is not possible due to insufficiently long piControl simulations, and also would not be applicable to a real world analysis where we do not know which of the models has internal variability most consistent with reality.

Figure 3 also shows results using informative prior information. Using additional prior information leads to more constrained results, which are closer to the true value of 1. Since our prior information [$\mathbf{P}(\boldsymbol{\beta})$, Eq. (11)] peaks at a scaling factor of 1 (see Fig. 2), this result is to be expected.

b. *Imperfect model study*

A perfect model setup might be expected to successfully estimate the correct forced response since the response in model fingerprints used should be identical to that in the pseudo observations, given that they all come from the same model. An imperfect model study now accounts for model error.

We use single historical simulations as pseudo observations and use our full analysis strategy (methods) to estimate a TCR value using model fingerprints taken from different models. The resulting estimate of TCR can then be compared to the known TCR value of our pseudo observations to evaluate how well our analysis is performing. Since each model is known to have different response patterns to different forcings, this analysis provides a test that is more relevant for the actual uncertainties than the perfect model results described in the previous section, as we do not know which model is most realistic compared to observations. It should be noted that this analysis will only be relevant to a real analysis with actual observations, if we can assume that ''models are statistically indistinguishable from the truth.'' It should also be noted that many of the models are not truly independent of each other, with different models sharing common ideas or code (Knutti et al. 2013). The models used as pseudo observations are listed in Tables 1 and 2.

Results for when one of these simulations, by the model ACCESS1.3, is used for the pseudo observations are shown as an illustration of the method (see Fig. 4). The fingerprints from different models give very different estimated TCR values. Similar to the perfect model test, the addition of more spatial and seasonal information leads to more constrained results, with the confidence interval decreasing for the individual model results. Crucially, it is not clear that the added diagnostics actually improve the chances of the results being accurate, and quite often they can lead to overconfidence. This seems particularly true for the case where the land–ocean contrast is included, where the results using fingerprints from different models are often quite well constrained but each set of TCR estimates differ greatly from each other, with nonoverlapping confidence intervals not always encompassing the correct answer (Fig. 4).

To calculate a combined probability for the imperfect model study, we integrate across the information from the individual model results by taking a weighted mean following Eq. (12), where each model is weighted equally (see section 3; this method will be subsequently referred to as the ''combined'' model analysis). Here this gives a wide uncertainty range, which in all cases also importantly encompasses the true TCR value (purple line; Fig. 4). The multimodel mean results (black line; Fig. 4) are very well constrained and give confidence intervals, which also include the correct answer in this case.

The results for all models and for our choices of four temporal and spatial climate change diagnostics are shown in Fig. 5. Our analysis method clearly has some skill, as in general higher TCR values are found for models used as pseudo observations whose true TCR

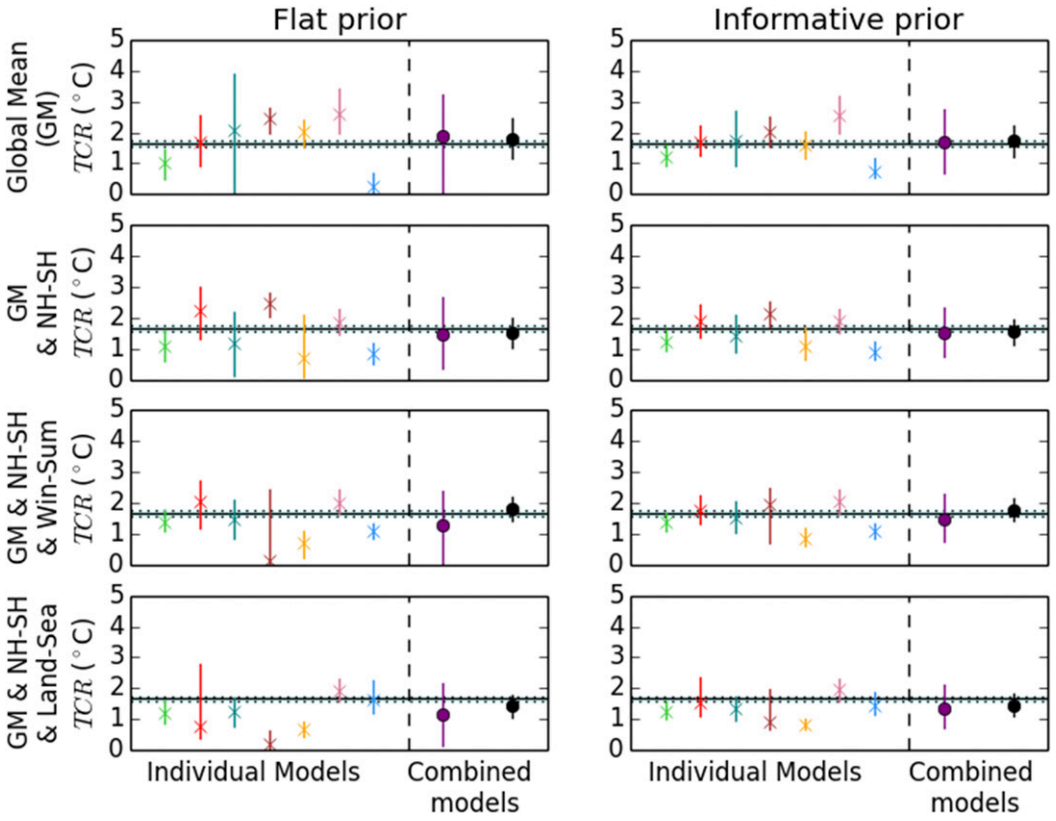

FIG. 4. The 5%–95% confidence intervals for TCR for the case where the ACCESS1.3 model is used as observations and model fingerprints are taken from a range of different models. See Table 1 for information on individual model colors; combined model results are in purple, and the multimodel mean is the black line. The vertical lines show the 5%–95% range, the symbols show median values, and the horizontal line shows the true TCR value.

values are higher and vice versa in both the combined individual model analysis and the results for the multimodel analysis, as shown by positive trend lines (null hypothesis for zero slope rejected in all panels at $p < 1 \times 10^{-4}$).

For the combined results (purple lines), for the majority of the pseudo observations, the true value lies within the 5%–95% range of our estimate. In fact, our analysis gives conservative confidence intervals (meaning that the 5%–95% confidence interval includes the real result in greater than the expected 90% of cases) in most of the analyses. Although in the perfect model analysis the additional diagnostics led to results closer to the truth, the opposite is the case in the imperfect model analysis, with the global mean on its own and in combination with the hemispheric mean proving to be most accurate and have less frequently wrong confidence intervals, than the more complex diagnostics. In particular, adding the seasonal difference causes an underestimate of the true TCR value with best estimates of TCR further from the truth. Using prior information on the scaling factors leads to improved performance, with more constrained results, lower distances from the best estimate to the truth, and fewer 5%–95% scaling ranges not containing the truth. It is noticeable that in all cases in Fig. 5 the informative prior leads to a reduced trend line through the best estimated TCR values, implying that when the actual TCR is greater the analysis will likely lead to an underestimate and when the actual TCR is lower the analysis will overestimate the result. This is due to the informative prior putting most weight on scaling factors closer to 1.

Using the multimodel mean (black lines) rather than a combination of individual models (which are calculated from smaller ensembles) gives much more constrained results. Although the distance from best estimate to truth for the global mean analysis is slightly larger than for the analysis using the combined models, for all other metrics the multimodel mean outperforms the combined results in terms of accuracy. This is because, unlike in the combined analysis, for the multimodel mean adding further diagnostics improves the performance considerably. Importantly though in many cases, the true value now lies outside the calculated confidence intervals. This provides evidence that results using the

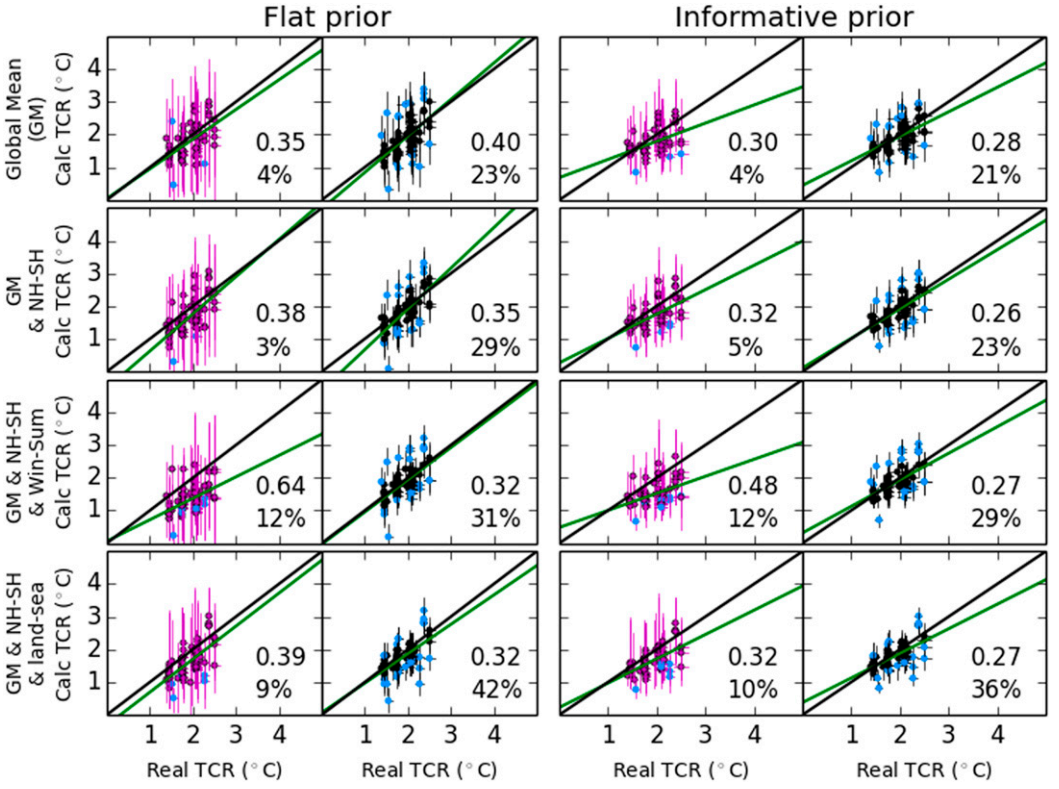

FIG. 5. Imperfect model results. Calculated TCR values plotted against actual TCR values for each of the 77 different models used here as observations, integrating over the individual model results (purple) and using the multimodel mean fingerprints (black). The 5%–95% confidence interval is shown by vertical lines. Best-guess estimate is shown by a circle: the circle is blue if the 5%–95% range does not encompasses the true value. The solid black line indicates where the truth lies; the solid green line shows the ordinary least squares regression through the best-guess estimates. The top number in each plot is the mean absolute error between the best-guess estimate and the truth; the second is the number of cases where the 5%–95% range does not include 1.

multimodel mean are overconfident, most likely because the multimodel mean analysis does not account for model uncertainty. For an estimate of TCR to be useful to inform policy decisions, it is the uncertainty range as much as the most likely value that is of importance. An overconfident estimate, such as that calculated by the multimodel mean analysis, should therefore be treated with caution. Given that the TCR range is already well constrained, using an informative prior has much less effect on the confidence intervals calculated for the multimodel analysis.

While some models appear to perform better than others, no model fingerprint proves consistently better than any other (Fig. S4). In addition, no individual model outperforms either the multimodel mean or the combined model results shown in Fig. 5.

Gillett et al. (2013) showed that there is uncertainty in the assumption that attributable warming scales linearly with TCR, finding considerable scatter around this relationship in model simulations. To test whether this, and not the difference in model response to external forcing, is the cause of our overconfident imperfect model results, we repeat our analysis, estimating attributable GHG warming instead of TCR, in individual model simulations where the actual value can be calculated directly. We define GHG warming as the linear annual trend in the global mean SAT in simulations forced with GHGs only during the period 1861–2005. Note that this analysis requires models to have run simulations forced by just GHG. Since not all models have these simulations, this analysis uses slightly fewer models that the TCR analysis. The results are shown in Fig. 6 and show that the true magnitude of attributable warming in a model's pseudo observation is similarly misestimated as the TCR, suggesting that the cause for the multimodel mean's overconfidence is not the TCR uncertainty but rather the neglect of model uncertainty in the attribution analysis. This also shows that using multimodel mean data for estimating attributable warming may also yield overconfident results.

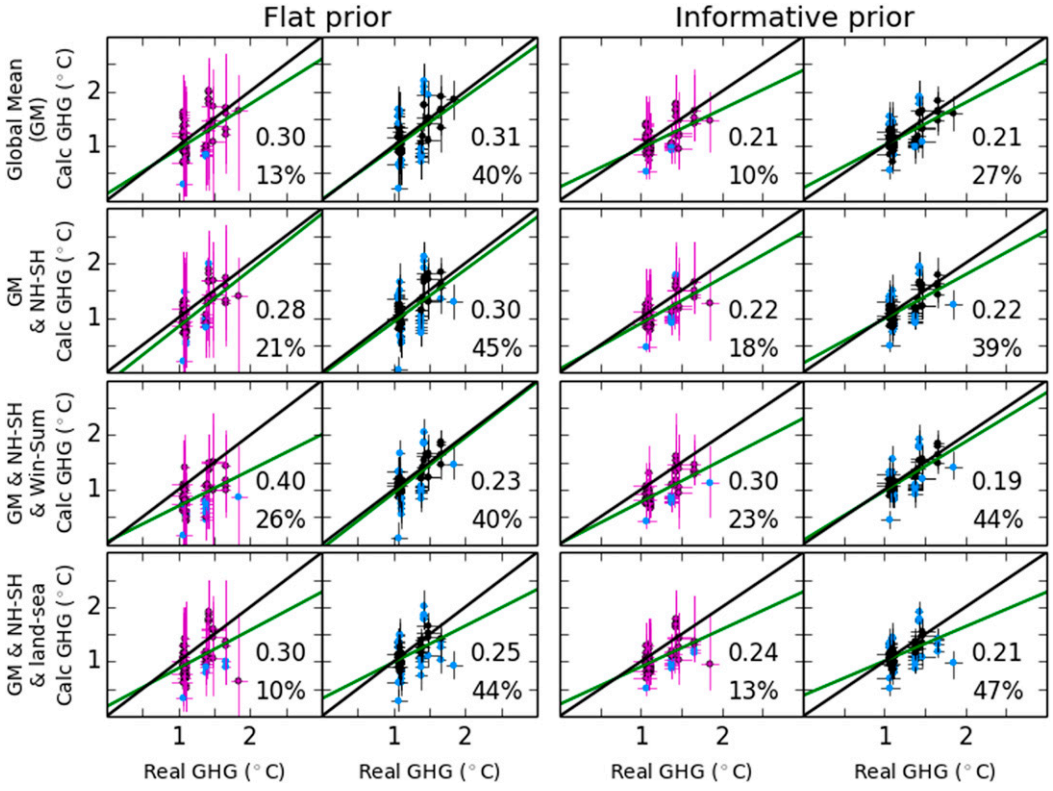

FIG. 6. Imperfect model results for attributable GHG warming. Attributed GHG trends plotted against actual GHG trends for 62 different models used as observations, integrating over the individual model results (purple) and using the multimodel mean fingerprints (black). Results are shown in the same format as for Fig. 5.

To address the overconfidence issue, we have repeated the multimodel mean analysis with a covariance matrix [$\mathbf{\Sigma}$ in Eq. (3)] formed from piControl simulations with more variance than in the initial simulations to include the unaccounted sources of uncertainty (such as differences in model response to forcing). This is a somewhat arbitrary treatment of uncertainties, but one which has been previously used for precipitation studies in order to address evidence of underestimated precipitation variability in models (e.g., Polson et al. 2013; Zhang et al. 2007). To calculate the most sensible factor to inflate the variance by, we repeat the imperfect model study with a range of covariance matrices (Fig. S2) calculated from piControl samples multiplied by different factors. We then choose the covariance matrices, for each of our spatial domains, which give failure rates close to what should be expected (failures outside the 90% confidence interval equal to 10%). Suitable factors for inflating the variance in the control simulations are found to range from 2.4 to 2.8. If we assume that models are statistically indistinguishable from the real world, then we can expect that an analysis using these covariance matrices would give more realistic confidence intervals.

5. Estimate of TCR from observations

We now use our method to provide an estimate of TCR constrained by the HadCRUT4 observational dataset. As described in section 2, there are 100 observational ensemble members which span the observational uncertainty range. First, we calculate results using only the median of these 100 (Fig. 7). Similar to the imperfect model results, the TCR estimates vary considerably when different models are used as fingerprints. This is in common with previous analyses (Gillett et al. 2013; Jones et al. 2013) that also found that different models give very different answers when estimating the magnitude of GHG warming and TCR. This difference is reduced considerably when an informative prior is used. When combined into a single probability distribution (purple line), the most constrained results are obtained when the hemispheric difference is used with the seasonal contrast, with results similar to the likely (≥66%) IPCC range. On the other hand, including the land–ocean contrast (in addition to the global mean and hemispheric difference) leads to large differences between the individual model results with nonoverlapping TCR estimate ranges, even when using the informative

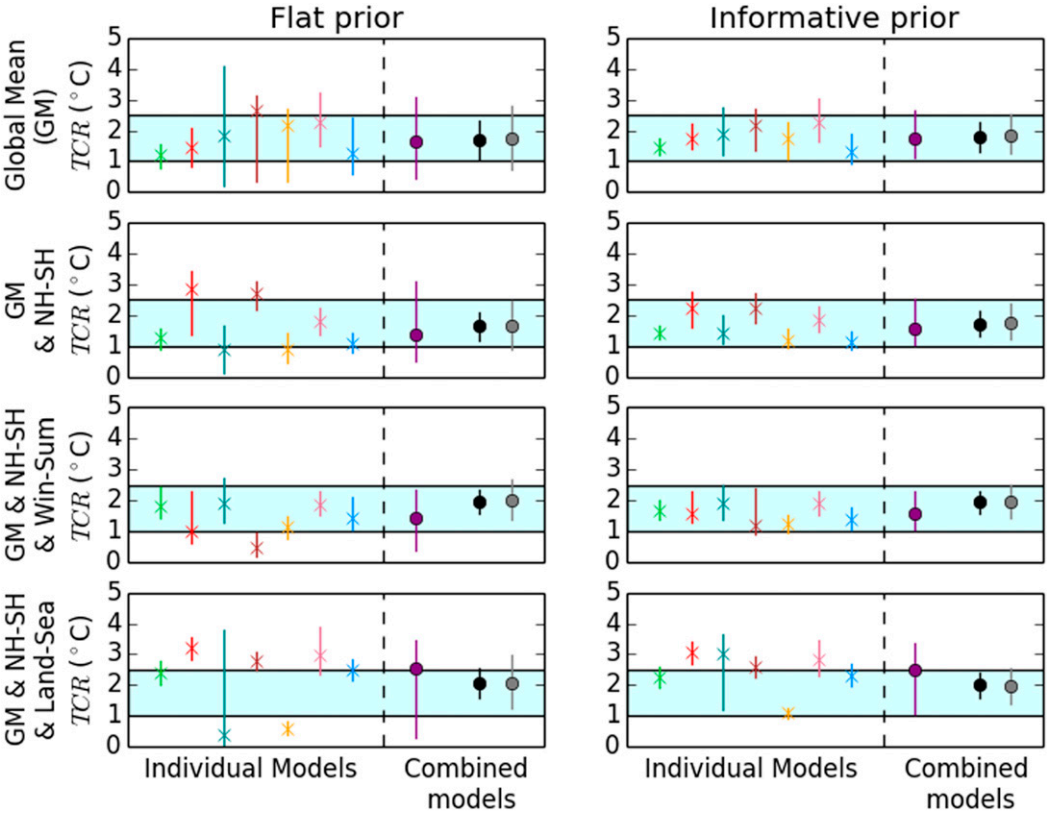

FIG. 7. TCR estimates from median observations and individual models (see Table 1). Combined results are in purple, the multimodel mean in black, and the multimodel mean with double piControl variance in gray. Ranges are 5%–95% probability. IPCC "likely" (i.e., $P > 0.66$) TCR range (1°–2.5°C) is shown by the light blue bar.

prior. The multimodel analysis gives much more constrained results (which are all within the IPCC range); in common with the imperfect model study, the informative prior has little effect on the final estimated TCR values. Also included in Fig. 7 are the multimodel mean results with inflated variance (plotted in gray). As expected, these are similar to the regular multimodel mean TCR estimates, except with wider confidence intervals.

To determine the sensitivity of these results to observational uncertainty, we repeat this analysis with each of the 100 separate observational ensemble members. The confidence intervals for the combined model results are shown for all 100 observational ensemble members in Fig. 8. The global mean results show very little sensitivity to the observational ensemble member and therefore observational uncertainty. This uncertainty increases when the hemispheric difference and seasonal difference are also used. When the land–ocean contrast is included in the analysis, results for different observational ensembles show large variations, and given the large observational uncertainty in this metric (see Fig. 1g), this is understandable.

As in Jones and Kennedy (2017), to obtain a final probability distribution for each of our chosen diagnostics, results from the 100 observational ensemble members are combined into a probability distribution by adding all of the individual probability distribution functions (pdfs), shown in Fig. 8, and normalizing; this is equivalent to Bayesian model averaging with a uniform prior. This combined pdf is shown in Fig. 9 for each of the combinations of climate diagnostics for the uninformative and informative priors and for the combined model analysis, and the multimodel mean analysis with and without inflated variance.

By comparing the confidence interval of the distribution integrated across all the observational ensemble members with that calculated from the median observations (Fig. 7), we can determine the importance of accounting for the observational uncertainty. For the multimodel mean analysis with the noninformative prior, the inclusion of observational uncertainty gives an increase of 11% in the 5%–95% confidence interval for the global mean metric, increasing to 30% for the metric that includes the land–sea contrast. This is of a similar order to that found by Jones and Kennedy (2017), who

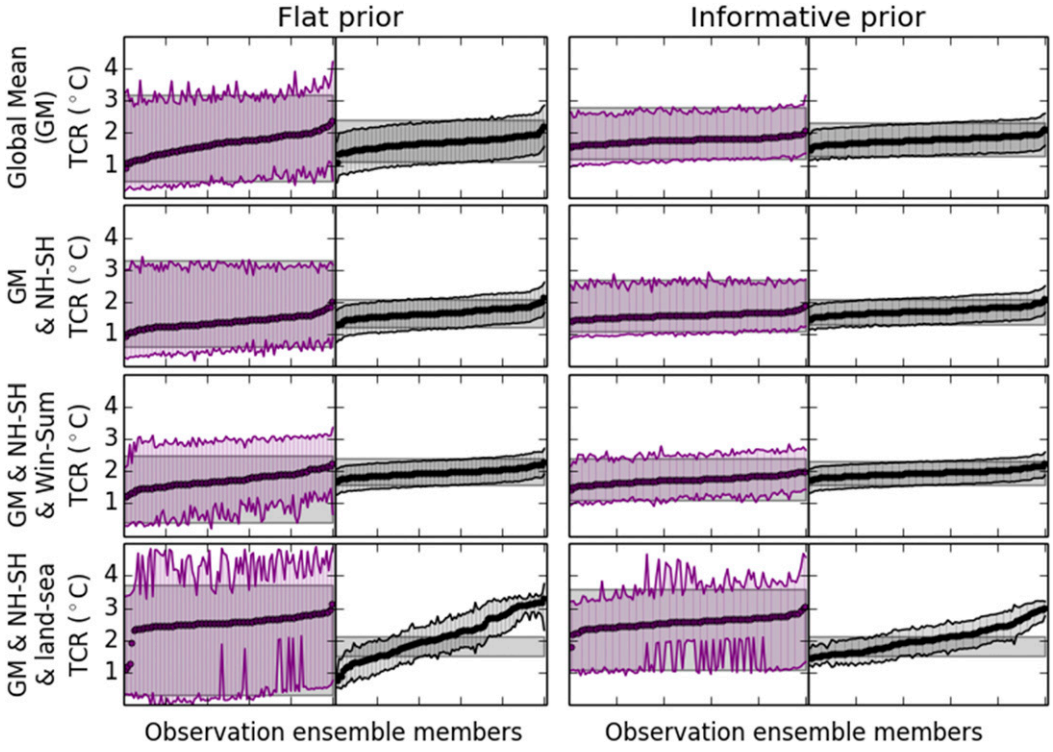

FIG. 8. Sensitivity of TCR estimate to observational uncertainty. Results shown for each of the 100 ensemble members, for combined model analysis (purple) and multimodel mean (black). The 5%–95% range for each ensemble member is shown by thin vertical lines, and they are linked by a solid line; the median probability is shown by a circle. The 5%–95% range calculated using the median observations (see Fig. 7) is shown by the gray shaded region. Ensemble members are plotted in ascending order for the median TCR value.

calculated an 18% increase in the 5%–95% range of the greenhouse gas scaling factor when accounting for observational uncertainty, in a similar detection and attribution analysis using multimodel mean spatiotemporal fingerprints.

Of the four different combinations of climate change diagnostics analyzed, the perfect model test suggests that including additional diagnostics, such as the hemispheric difference, land–sea contrast and seasonal contrast, should in theory lead to better constrained TCR estimates. Our imperfect study, however, suggests that the additional diagnostics do not improve the results but instead increase the distance between the best estimate and the truth. Similarly, TCR estimates are improved in the imperfect model analysis by using an informative prior (smaller error between best estimate and truth, and more confidence intervals containing the truth). Consequently, this points to the analysis with just the global mean or with the global mean and the hemispheric contrast in combination with an informative prior giving the most reliable results. Although it should be noted that, as Fig. 5 shows, because the informative prior favors TCR values closer to the model values, if the true TCR value is substantially higher or lower than the model values, this choice of prior could lead to unreliable estimates The TCR confidence interval (90% range) for this preferred diagnostic (global mean plus hemispheric contrast) is 0.4°–3.1°C, which reduces to 1.0°–2.6°C (best estimate 1.6°C), when using an informative prior (see Fig. 9). This spans the likely (≥66%) IPCC range (1°–2.5°C; note, however, that the IPCC range is using uncertainty ranges that are increased by expert assessment in order to account for structural uncertainty and unknown unknowns, and hence cannot be readily compared to a 66% range arising from a statistical analysis). The imperfect model study suggests that our estimated TCR range could be overly conservative.

The multimodel mean, on the other hand, is much more constrained in our results with the regular covariance, consistent with the IPCC estimate for all of the analysis choices. For the set of diagnostics with global mean and hemispheric contrast, the 90% range of TCR was found to be 1.1°–2.2°C with a best estimate of 1.7°C, which reduces to 1.3°–2.2°C with an informative prior. The imperfect model study did find that the multimodel results were overconfident. This was corrected for by calculating covariance matrices with inflated variance.

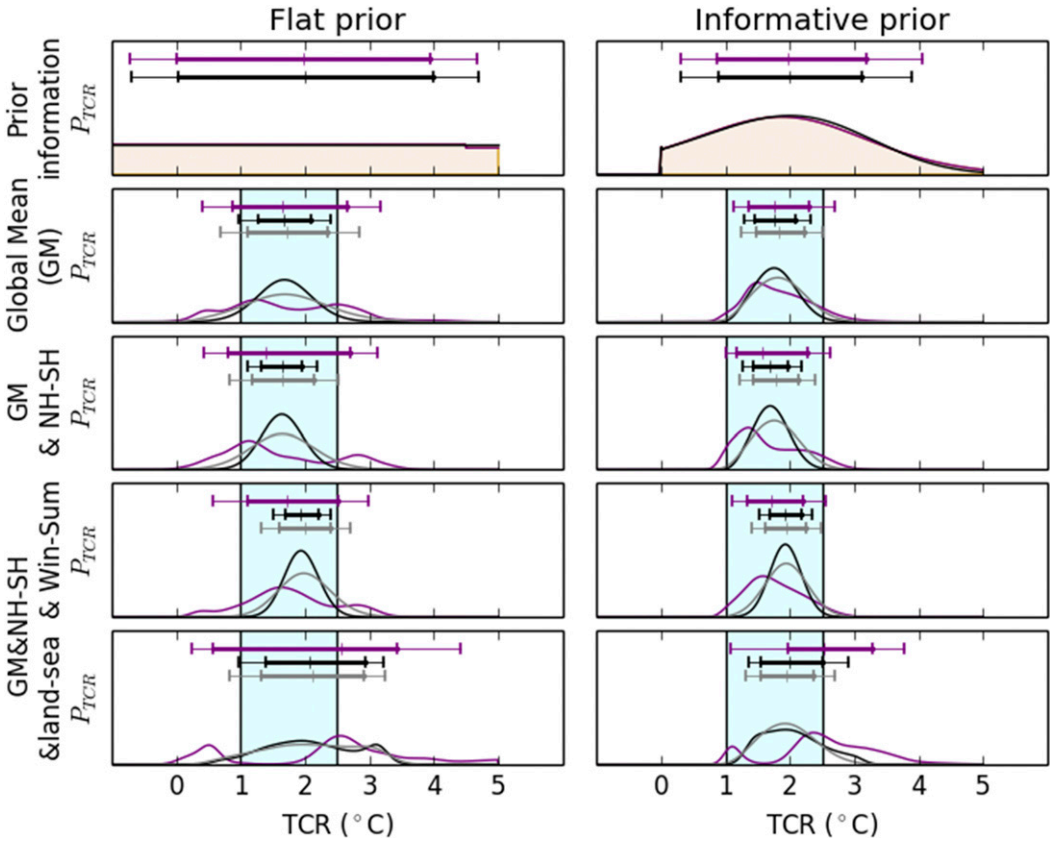

FIG. 9. TCR pdfs for all observational ensemble members and individual model fingerprints. Distributions show combined pdf results for the combined model results (purple), multimodel mean (black), and multimodel mean with inflated variance (gray) with horizontal bars giving the 5%–95% and 17%–83% range. The prior range is shown by the orange shading (note prior amplitude is arbitrary and has been rescaled for illustrative purposes). IPCC "likely" (i.e., ≥0.66) TCR range (1°–2.5°C) is shown by the light blue bar.

In the multimodel analysis using the global mean and hemispheric contrast a factor of 2.6 was found to be most suitable by the imperfect model study. Using this covariance matrix for the analysis with the informative prior information, a 90% range of TCR of 1.2°–2.4°C with a best estimate of 1.8°C is estimated, which is slightly wider than the raw multimodel TCR estimate, but still more tightly constrained than the integrated TCR estimate.

It is interesting to compare our TCR range with those from similar analyses by other authors. Gillett et al. (2013), using an EIV method accounting for model uncertainty, calculated a 5%–95% range of 0.9°–2.3°C, and Jones et al. (2016) calculated a 5%–95% range of 1.1°–2.1°C, both using the multimodel mean. These two confidence intervals are very similar to our multimodel results using the global mean and hemispheric contrast without an informative prior and using the standard covariance matrix (not inflated variance), which gives a 5%–95% confidence interval of 1.1°–2.2°C.

As described in section 2, unlike Gillett et al. (2013) and Jones et al. (2013, 2016), all our results have been obtained using blended temperatures using the method described in Cowtan et al. (2015). This is known to reduce warming in the models and would therefore be expected to increase our estimate of TCR. To determine the sensitivity of our results to this compared to the more usual use of just surface air temperatures, we repeated the analysis for the median of the HadCRUT4 observational ensemble, but this time using just SATs to calculate the model fields. The results are shown in Fig. S3 in the supplementary material. Using only SATs does indeed result in lower values for TCR, although the difference is relatively small (varying between 3% and 5% lowered estimate of TCR for the multimodel mean analysis). This is a smaller effect than reported by Richardson et al. (2016), who found that using only SATs reduced the estimated TCR value by 7%–9%; however, this value is for an analysis using model data with observations with perfect coverage, whereas in this

study we have used observations with missing data and have masked our model data to the same coverage.

In this study we have used full-coverage global SATs from 1PCT simulations to calculate the model's TCR used in Eq. (12) (see section 2). This is the standard metric used, for example, by the IPCC (Collins et al. 2013). We could instead have calculated a model's TCR using a blend of SATs and SSTs in 1PCT simulations. This is likely to have led to reduced values for **P**(TCR) [Eq. (12); see, e.g., Cowtan et al. 2015], which would lead to an equivalent reduction in our observationally constrained TCR values. Our final TCR estimate would be likely reduced even further if we had additionally decided to calculate TCR from 1PCT simulations masked to observational coverage (Cowtan et al. 2015). Which definition of TCR to use depends on what measure of global mean surface temperature rise we are interested in (see, e.g., Schurer et al. 2018); we have decided to use the definition we have, so that our results are directly comparable with previous published estimates (Knutti et al. 2017).

6. Conclusions

In this study, we have implemented an adapted optimal detection analysis within a Bayesian framework to derive estimates for TCR. We have tested our analysis approach with a perfect and imperfect model study and have shown that our method can correctly estimate TCR values using fingerprints taken from different models. We have evaluated the use of different climate diagnostics and the effect of combining results from individual models versus using the multimodel mean fingerprint.

The best estimate from the combined model analysis is 1.7°C with a 90% range of 1.0°–2.6°C, which method tests indicate is likely to be conservative. The best estimate from the multimodel analysis is 1.7°C with a 90% range of 1.3°–2.2°C, but this has been shown to be overconfident in the presence of model uncertainty. To compensate for this we have carried out an alternate multimodel mean analysis with increased variance in the covariance matrix; this gives a 90% range of 1.2°–2.4°C. These three strands of evidence combined strongly supports a true TCR estimate lying within the central part of the IPCC estimated range and rules out very low or very high values, although this is, at least partly, due to the choice of prior used.

In common with Jones and Kennedy (2017), observational uncertainty has a moderate effect on the estimated TCR range, increasing the uncertainty in TCR by 11%–30% depending on the details of the analysis, with the land–sea contrast leading to the highest increase. Representing the model fingerprints as a blend of SATs and SSTs, rather than just SATs as is commonly done, has a relatively small role on the TCR values increasing the estimate by about 3%–5%.

This study also reaches a number of other findings of interest to the detection and attribution community. The perfect model results highlight the importance of internal variability learnt from the piControl experiments for estimating the confidence interval on scaling factors. In cases where we know that the internal variability of a model is less than the average of the control samples used, our estimates prove to be conservative, and in the cases where the internal variability is larger our results can be far too overconfident. This should motivate the estimation of the actual internal variability of the climate and the use of models with realistic internal variability when estimating uncertainty ranges.

The imperfect model results raise interesting questions about the use of more spatial and seasonal information in the analysis, with more complex diagnostics leading to less robust results for many models. The multimodel mean fingerprints, which are widely used, have been shown to give more robust results both for estimates of TCR and attributable warming than any individual model, but, using the standard detection and attribution framework (which does not account for model uncertainty), gives overconfident TCR estimates. We would therefore recommend caution in the interpretation of detection and attribution analyses using just the multimodel mean fingerprints, which do not account for model uncertainty. Despite the poor performance by individual model results on their own, integrating of these individual results into a combined distribution yields a far more reliable result, with the best estimate closer to the real value than for any individual model. In the analysis with just the global mean, the combined distribution performs better at estimating a best-fit value than use of the multimodel mean fingerprint, which suggests that the combination of individual model results in this manner has potential and certainly lends confidence to our final results.

The use of prior information has been shown to lead to more constrained results, and further studies could investigate including different prior information into the analysis. For example, a constraint on the anthropogenic aerosol forcing that makes use of the spatial pattern of aerosol forcing evaluated against data would be extremely valuable to break the degeneracy with the GHG forcing (Malavelle et al. 2017).

Acknowledgments. We acknowledge Kevin Cowtan for making his code and results available and for help in their use and Francis Zwiers and Laura Wilcox for helpful comments. We thank two anonymous reviewers for their insightful comments that helped improve the

manuscript. A.S., G.H., S.T., and C.M. were supported by the ERC funded project TITAN (EC-320691); A.S. and G.H. were supported by NERC under the Belmont forum, Grant PacMedy (NE/P006752/1); G.H. and S.T. were supported by NCAS (R8/H12/83/029); C.M. was supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra; and G.H. was further funded by the Wolfson Foundation and the Royal Society as a Royal Society Wolfson Research Merit Award (WM130060) holder and by the NERC-funded SMURPHS project. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, the climate modeling groups for producing and making available their model output, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison, and the Global Organization for Earth System Science Portals for Earth System Science Portals.

REFERENCES

Allen, M. R., and S. F. B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Climate Dyn.*, **15**, 419–434, https://doi.org/10.1007/s003820050291.

——, and P. A. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting, part I: Theory. *Climate Dyn.*, **21**, 477–491, https://doi.org/10.1007/s00382-003-0313-9.

——, D. J. Frame, C. Huntingford, C. D. Jones, J. A. Lowe, M. Meinshausen, and N. Meinshausen, 2009: Warming caused by cumulative carbon emissions towards the trillionth tonne. *Nature*, **458**, 1163–1166, https://doi.org/10.1038/nature08019.

Berliner, L. M., R. A. Levine, and D. J. Shea, 2000: Bayesian climate change assessment. *J. Climate*, **13**, 3805–3820, https://doi.org/10.1175/1520-0442(2000)013<3805:BCCA>2.0.CO;2.

Bindoff, N. L., and Coauthors, 2013: Detection and attribution of climate change: From global to regional. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 867–952.

Collins, M., and Coauthors, 2013: Long-term climate change: Projections, commitments and irreversibility. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 1029–1136.

Cowtan, K., and Coauthors, 2015: Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophys. Res. Lett.*, **42**, 6526–6534, https://doi.org/10.1002/2015GL064888.

Frame, D. J., D. A. Stone, P. A. Stott, and M. R. Allen, 2006: Alternatives to stabilization scenarios. *Geophys. Res. Lett.*, **33**, L14707, https://doi.org/10.1029/2006GL025801.

Gillett, N. P., V. K. Arora, D. Matthews, and M. R. Allen, 2013: Constraining the ratio of global warming to cumulative CO_2 emissions using CMIP5 simulations. *J. Climate*, **26**, 6844–6858, https://doi.org/10.1175/JCLI-D-12-00476.1.

Hannart, A., A. Ribes, and P. Naveau, 2014: Optimal fingerprinting under multiple sources of uncertainty. *Geophys. Res. Lett.*, **41**, 1261–1268, https://doi.org/10.1002/2013GL058653.

Hasselmann, K., 1993: Optimal fingerprints for the detection of time-dependent climate change. *J. Climate*, **6**, 1957–1971, https://doi.org/10.1175/1520-0442(1993)006<1957:OFFTDO>2.0.CO;2.

Hegerl, G. C., and G. R. North, 1997: Comparison of statistically optimal approaches to detecting anthropogenic climate change. *J. Climate*, **10**, 1125–1133, https://doi.org/10.1175/1520-0442(1997)010<1125:COSOAT>2.0.CO;2.

——, and M. R. Allen, 2002: Origins of model–data discrepancies in optimal fingerprinting. *J. Climate*, **15**, 1348–1356, https://doi.org/10.1175/1520-0442(2002)015<1348:OOMDDI>2.0.CO;2.

——, and F. Zwiers, 2011: Use of models in detection and attribution of climate change. *Wiley Interdiscip. Rev.: Climate Change*, **2**, 570–591, https://doi.org/10.1002/wcc.121.

——, K. Hasselmann, U. Cubasch, J. F. B. Mitchell, E. Roeckner, R. Voss, and J. Waszkewitz, 1997: Multi-fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas-plus-aerosol and solar forced climate change. *Climate Dyn.*, **13**, 613–634, https://doi.org/10.1007/s003820050186.

——, T. J. Crowley, W. T. Hyde, and D. J. Frame, 2006: Climate sensitivity constrained by temperature reconstructions over the past seven centuries. *Nature*, **440**, 1029–1032, https://doi.org/10.1038/nature04679.

——, and Coauthors, 2007: Understanding and attributing climate change. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 663–745.

Huntingford, C., P. A. Stott, M. R. Allen, and F. H. Lambert, 2006: Incorporating model uncertainty into attribution of observed temperature change. *Geophys. Res. Lett.*, **33**, L05710, https://doi.org/10.1029/2005GL024831.

Jones, G. S., and J. J. Kennedy, 2017: Sensitivity of attribution of anthropogenic near-surface warming to observational uncertainty. *J. Climate*, **30**, 4677–4691, https://doi.org/10.1175/JCLI-D-16-0628.1.

——, P. A. Stott, and N. Christidis, 2013: Attribution of observed historical near-surface temperature variations to anthropogenic and natural causes using CMIP5 simulations. *J. Geophys. Res. Atmos.*, **118**, 4001–4024, https://doi.org/10.1002/jgrd.50239.

——, ——, and J. F. B. Mitchell, 2016: Uncertainties in the attribution of greenhouse gas warming and implications for climate prediction. *J. Geophys. Res. Atmos.*, **121**, 6969–6992, https://doi.org/10.1002/2015JD024337.

Kent, E. C., and Coauthors, 2017: A call for new approaches to quantifying biases in observations of sea surface temperature. *Bull. Amer. Meteor. Soc.*, **98**, 1601–1616, https://doi.org/10.1175/BAMS-D-15-00251.1.

Kettleborough, J. A., B. B. B. Booth, P. A. Stott, and M. R. Allen, 2007: Estimates of uncertainty in predictions of global mean surface temperature. *J. Climate*, **20**, 843–855, https://doi.org/10.1175/JCLI4012.1.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G.A. Meehl, 2010: Challenges in combining projections from multiple climate models. *J. Climate*, **23**, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1.

——, D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophys. Res. Lett.*, **40**, 1194–1199, https://doi.org/10.1002/grl.50256.

——, M. A. A. Rugenstein, and G. C. Hegerl, 2017: Beyond equilibrium climate sensitivity. *Nat. Geosci.*, **10**, 727–736, https://doi.org/10.1038/ngeo3017.

Lee, T. C. K., F.W. Zwiers, G.C. Hegerl, X. Zhang, and M. Tsao, 2005: A Bayesian climate change detection and attribution assessment. *J. Climate*, **18**, 2429–2440, https://doi.org/10.1175/JCLI3402.1.

Malavelle, F. F., and Coauthors, 2017: Strong constraints on aerosol–cloud interactions from volcanic eruptions. *Nature*, **546**, 485–491, https://doi.org/10.1038/nature22974.

Matthews, H. D., N. P. Gillett, P. A. Stott, and K. Zickfeld, 2009: The proportionality of global warming to cumulative carbon emissions. *Nature*, **459**, 829–832, https://doi.org/10.1038/nature08047.

Ming, Y., and V. Ramaswamy, 2009: Nonlinear climate and hydrological responses to aerosol effects. *J. Climate*, **22**, 1329–1339, https://doi.org/10.1175/2008JCLI2362.1.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, https://doi.org/10.1029/2011JD017187.

Myhre, G., and et al, 2013: Anthropogenic and natural radiative forcing. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 659–740.

Polson, D., G. C. Hegerl, X. Zhang, T. J. Osborn, D. Polson, G. C. Hegerl, X. Zhang, and T. J. Osborn, 2013: Causes of robust seasonal land precipitation changes. *J. Climate*, **26**, 6679–6697, https://doi.org/10.1175/JCLI-D-12-00474.1.

Ribes, A., and L. Terray, 2013: Application of regularised optimal fingerprinting to attribution. Part II: Application to global near-surface temperature. *Climate Dyn.*, **41**, 2837–2853, https://doi.org/10.1007/s00382-013-1736-6.

——, S. Planton, and L. Terray, 2013: Application of regularised optimal fingerprinting to attribution. Part I: Method, properties and idealised analysis. *Climate Dyn.*, **41**, 2817–2836, https://doi.org/10.1007/s00382-013-1735-7.

——, N. P. Gillett, F. W. Zwiers, A. Ribes, N. P. Gillett, and F. W. Zwiers, 2015: Designing detection and attribution simulations for CMIP6 to optimize the estimation of greenhouse gas–induced warming. *J. Climate*, **28**, 3435–3438, https://doi.org/10.1175/JCLI-D-14-00691.1.

——, F. W. Zwiers, J.-M. Azaïs, and P. Naveau, 2017: A new statistical approach to climate change detection and attribution. *Climate Dyn.*, **48**, 367–386, https://doi.org/10.1007/s00382-016-3079-6.

Richardson, M., K. Cowtan, E. Hawkins, and M. B. Stolpe, 2016: Reconciled climate response estimates from climate models and the energy budget of Earth. *Nat. Climate Change*, **6**, 931–935, https://doi.org/10.1038/nclimate3066.

Schnur, R., and K. l. Hasselmann, 2005: Optimal filtering for Bayesian detection and attribution of climate change. *Climate Dyn.*, **24**, 45–55, https://doi.org/10.1007/s00382-004-0456-3.

Schurer, A. P., K. Cowtan, E. Hawkins, M. E. Mann, V. Scott, and S. F. B. Tett, 2018: Interpretations of the Paris climate target. *Nat. Geosci.*, **11**, 220–221, https://doi.org/10.1038/s41561-018-0086-8.

Shindell, D. T., 2014: Inhomogeneous forcing and transient climate sensitivity. *Nat. Climate Change*, **4**, 274–277, https://doi.org/10.1038/nclimate2136.

Stevens, B., 2015: Rethinking the lower bound on aerosol radiative forcing. *J. Climate*, **28**, 4794–4819, https://doi.org/10.1175/JCLI-D-14-00656.1.

Stott, P. A., J. F. Mitchell, M. R. Allen, T. L. Delworth, J. M. Gregory, G. A. Meehl, and B. D. Santer, 2006: Observational constraints on past attributable warming and predictions of future global warming. *J. Climate*, **19**, 3055–3069, https://doi.org/10.1175/JCLI3802.1.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Tett, S. F. B., and Coauthors, 2002: Estimation of natural and anthropogenic contributions to twentieth century temperature change. *J. Geophys. Res.*, **107**, 4306, https://doi.org/10.1029/2000JD000028.

——, and Coauthors, 2007: The impact of natural and anthropogenic forcings on climate and hydrology since 1550. *Climate Dyn.*, **28**, 3–34, https://doi.org/10.1007/s00382-006-0165-1.

Wang, H., S.-P. Xie, Q. Liu, H. Wang, S.-P. Xie, and Q. Liu, 2016: Comparison of climate response to anthropogenic aerosol versus greenhouse gas forcing: Distinct patterns. *J. Climate*, **29**, 5175–5188, https://doi.org/10.1175/JCLI-D-16-0106.1.

Wilcox, L. J., E. J. Highwood, and N. J. Dunstone, 2013: The influence of anthropogenic aerosol on multi-decadal variations of historical global climate. *Environ. Res. Lett.*, **8**, 024033, https://doi.org/10.1088/1748-9326/8/2/024033.

Xie, S.-P., B. Lu, and B. Xiang, 2013: Similar spatial patterns of climate responses to aerosol and greenhouse gas changes. *Nat. Geosci.*, **6**, 828–832, https://doi.org/10.1038/ngeo1931.

Zhang, X., F. W. Zwiers, G. C. Hegerl, F. H. Lambert, N. P. Gillett, S. Solomon, P. A. Stott, and T. Nozawa, 2007: Detection of human influence on twentieth-century precipitation trends. *Nature*, **448**, 461–465, https://doi.org/10.1038/nature06025.